JHEP

Published for SISSA by Springer

Received: December 6, 2018
Revised: February 18, 2019
Accepted: April 18, 2019
Published: May 7, 2019

Variational autoencoders for new physics mining at the Large Hadron Collider

Olmo Cerri,[a] **Thong Q. Nguyen,**[a] **Maurizio Pierini,**[b] **Maria Spiropulu**[a] **and Jean-Roch Vlimant**[a]

[a] *California Institute of Technology, 1200 E California Blvd, Pasadena, CA 91125, U.S.A.*
[b] *CERN, Espl. des Particules 1, 1217 Meyrin, Switzerland*

E-mail: olmo@caltech.edu, thong@caltech.edu, Maurizio.Pierini@cern.ch, smaria@caltech.edu, jvlimant@caltech.edu

Abstract: Using variational autoencoders trained on known physics processes, we develop a one-sided threshold test to isolate previously unseen processes as outlier events. Since the autoencoder training does not depend on any specific new physics signature, the proposed procedure doesn't make specific assumptions on the nature of new physics. An event selection based on this algorithm would be complementary to classic LHC searches, typically based on model-dependent hypothesis testing. Such an algorithm would deliver a list of anomalous events, that the experimental collaborations could further scrutinize and even release as a catalog, similarly to what is typically done in other scientific domains. Event topologies repeating in this dataset could inspire new-physics model building and new experimental searches. Running in the trigger system of the LHC experiments, such an application could identify anomalous events that would be otherwise lost, extending the scientific reach of the LHC.

Keywords: Beyond Standard Model, Hadron-Hadron scattering (experiments)

ArXiv ePrint: 1811.10276

https://doi.org/10.1007/JHEP05(2019)036

JHEP05(2019)036

Contents

JHEP05(2019)036

1 Introduction

One of the main motivations behind the construction of the CERN Large Hadron Collider (LHC) is the exploration of the high-energy frontier in search for *new physics* phenomena. New physics could answer some of the standing fundamental questions in particle physics, e.g., the nature of dark matter or the origin of electroweak symmetry breaking. In LHC experiments, searches for physics beyond the Standard Model (BSM) are typically carried on as fully-supervised data analyses: assuming a new physics scenario of some kind, a search is structured as a hypothesis test, based on a profiled-likelihood ratio [1]. These searches are said to be *model dependent*, since they depend on considering a specific new physics model.

Assuming that one is testing the *right* model, this approach is very effective in discovering a signal, as demonstrated by the discovery of the Standard Model (SM) Higgs boson [2, 3] at the LHC. On the other hand, given the (so far) negative outcome of many BSM searches at particle-physics experiments, it is possible that a future BSM model, if any, is not among those typically tested. The problem is more profound if analyzed in the context of the LHC big-data problem: at the LHC, 40 million proton-beam collisions are produced every second, but only ∼1000 collision events/sec can be stored by the ATLAS and CMS experiments, due to limited bandwidth, processing, and storage resources. It is possible to imagine BSM scenarios that would escape detection, simply because the corresponding new physics events would be rejected by a typical set of online selection algorithms.

Establishing alternative search methodologies with reduced model dependence is an important aspect of future LHC runs. Traditionally, this issue was addressed with so-called model-independent searches, performed at the Tevatron [4, 5], at HERA [6], and at the LHC [7, 8], as discussed in section 2.

In this paper, we propose to address this need by deploying an unsupervised algorithm in the online selection system (trigger) of the LHC experiments.[1] This algorithm would be trained on known SM processes and could be able to identify BSM events as anomalies. The selected events could be stored in a special stream, scrutinized by experts (e.g., to exclude the occurrence of detector malfunctions that could explain the anomalies), and even released outside the experimental collaborations, in the form of an open-access catalog. The final goal of this application is to identify anomalous event topologies and inspire future supervised searches on data collected afterwards.

As an example, we consider the case of a typical single-lepton data stream, selected by a hardware-based Level-1 (L1) trigger system. In normal conditions, the L1 trigger is the first of a two-steps selection stage. After a coarse (and often local) reconstruction and loose selection at L1, events are fully reconstructed in the High Lever Trigger (HLT), where a much tighter selection is applied. The selection is usually done having in mind specific signal topologies, eg., specific BSM models. In this study, we imagine to replace this model-dependent selection with a variational autoencoder (VAE) [11, 12] looking for anomalous events in the incoming single-lepton stream. The VAE is trained to compress the input event representation into a lower-dimension latent space and then decompress it, returning the shape parameters describing the probability density function (pdf) of each input quantity given a point in the compressed space. In addition, a VAE allows a stochastic modeling of the latent space, a feature which is missing in a simple AE architecture. The highlighted procedure is not specific of the considered single-lepton stream and could be easily extended to other data streams.

The distribution of the VAE's reconstruction loss on a validation sample is used to define a threshold, corresponding to a desired acceptance rate for SM events. All the events with loss larger than the threshold are considered as potential anomalies and could be stored in a low-rate anomalous-event data stream. In this work, we set the threshold such that ~ 1000 SM events would be collected every month under typical LHC operation conditions. In particular, we took as a reference 8 months of data taking per year, with an integrated luminosity of $\sim 40\,\text{fb}^{-1}$. Assuming an LHC duty cycle of 2/3, this corresponds to an average instantaneous luminosity of $\sim 2.9 \times 10^{33}\ \text{cm}^{-2}\ \text{s}^{-1}$.

We then evaluate the BSM production cross section that would correspond to a signal excess of 100 BSM events selected per month, as well as the one that would give a signal yield $\sim 1/3$ of the SM yield. For this, we consider a set of low-mass BSM resonances, decaying to one or more leptons and light enough to be challenging for the currently employed LHC trigger algorithms.

[1] A description of the ATLAS and CMS trigger systems can be found in ref. [9] and ref. [10], respectively. In this study, we take the data-taking strategy of these two experiments as a reference. On the other hand, the proposed strategy could be adapted to other use cases.

JHEP05(2019)036

This paper is structured as follows: we discuss related works in section 2. Section 3 gives a brief description of the dataset used. Section 4 describes the VAE model used in the study, as well as a set of fully-supervised classifiers used for performance comparison. Results are discussed in section 5. In section 6 we discuss how such a procedure could be deployed in a typical LHC experiment while relying exclusively on data. Conclusions are given in section 7. Appendix A provides a brief comparison between VAEs and plain autoencoders (AEs).

2 Related work

Model-independent searches for new physics have been performed at the Tevatron [4, 5], HERA [6], and the LHC [7, 8]. These searches are based on the comparison of a large set of binned distributions to the prediction from Monte Carlo (MC) simulations, in search for bins exhibiting a deviation larger than some predefined threshold. While the effectiveness of this strategy in establishing a discovery has been a matter of discussion, a recent study by the ATLAS collaboration [8] rephrased this model-independent search strategy into a tool to identify interesting excesses, on which traditional analysis techniques could be performed on independent datasets (e.g., the data collected after running the model-independent analysis). This change of scope has the advantage of reducing the trial factor (i.e., the so-called *look-elsewhere* effect [13, 14]), which would otherwise wash out the significance of an observed excess.

Our strategy is similar to what is proposed in ref. [8], with two substantial differences: (i) we aim to process also those events that could be discarded by the online selection, by running the algorithm as part of the trigger process; (ii) we do so exploiting deep-learning-based anomaly detection techniques.

Applying deep learning at the trigger level has been proposed in ref. [15]. Recent works [16–19] have investigated the use of machine-learning techniques to setup new strategies for BSM searches with minimal or no assumption on the specific new-physics scenario under investigation. In this work, we use VAEs [11, 12] based on high-level features as a baseline. Previously, autoencoders have been used in collider physics for detector monitoring [20, 21] and event generation [22]. Autoencoders have also been explored to define a jet tagger that would identify new physics events with anomalous jets [23, 24], with a strategy similar to what we apply to the full event in this work.

Anomaly detection has been a traditional use case for one-class machine learning methods, such as one-class Support Vector Machine [25] or Isolation Forest [26, 27]. A review of proposed methods can be found in ref. [28]. Variational methods have been shown to be effective for novelty detection, as for instance is discussed in ref. [29]. In particular, VAEs [11] have been proposed as an effective method for anomaly detection [12].

3 Data samples

The dataset used for this study is a refined version of the high-level-feature (HLF) dataset used in ref. [15]. Proton-proton collisions are generated using the `PYTHIA8` event-generation library [30], fixing the center-of-mass energy to the LHC Run-II value (13 TeV) and the

JHEP05(2019)036

average number of overlapping collisions per beam crossing (pileup) to 20. These beam conditions loosely correspond to the LHC operating conditions in 2016.

Events generated by PYTHIA8 are processed with the DELPHES library [31], to emulate detector efficiency and resolution effects. We take as a benchmark detector description the upgraded design of the CMS detector, foreseen for the High-Luminosity LHC phase [32]. In particular, we use the CMS HL-LHC detector card distributed with DELPHES. We run the DELPHES *particle-flow* (PF) algorithm, which combines the information from different detector components to derive a list of reconstructed particles, the so-called PF candidates. For each particle, the algorithm returns the measured energy and flight direction. Each particle is associated to one of three classes: charged particles, photons, and neutral hadrons. In addition, lists of reconstructed electrons and muons are given.

Many SM processes would contribute to the considered single-lepton dataset. For simplicity, we restrict the list of relevant SM processes to the four with the highest production cross sections, namely:

- Inclusive W production, with $W \to \ell\nu$ ($\ell = e,\ \mu,\ \tau$).
- Inclusive Z production, with $Z \to \ell\ell$ ($\ell = e,\ \mu,\ \tau$).
- $t\bar{t}$ production.
- QCD multijet production.[2]

These samples are mixed to provide a SM cocktail dataset, which is then used to train autoencoder models and to tune the threshold requirement that defines what we consider an anomaly. The cocktail is built scaling down the high-statistics samples ($t\bar{t}$, W, and Z) to the lowest-statistics one (QCD, whose generation is the most computing-expensive), according to their production cross-section values (estimated at leading order with PYTHIA) and selection efficiencies, shown in table 1.

Events are filtered at generation requiring an electron, muon, or tau lepton with $p_T > 22\,\text{GeV}$. Once detector effects are taken into account through the DELPHES simulation, events are further selected requiring the presence of one reconstructed lepton (electron or muon) with transverse momentum $p_T > 23\,\text{GeV}$ and a loose isolation requirement $\text{Iso} < 0.45$. If more than one reconstructed lepton is present, the highest p_T one is considered. The isolation for the considered lepton ℓ is computed as:

$$\text{Iso} = \frac{\sum_{p \neq \ell} p_T^p}{p_T^\ell}\,, \tag{3.1}$$

where the index p runs over all the photons, charged particles, and neutral hadrons within a cone of size $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2} < 0.3$ from ℓ.[3]

[2] To speed up the generation process for QCD events, we require $\sqrt{\hat{s}} > 10\,\text{GeV}$, the fraction of QCD events with $\sqrt{\hat{s}} < 10\,\text{GeV}$ and producing a lepton within acceptance being negligible but computationally expensive.

[3] As common for collider physics, we use a Cartesian coordinate system with the z axis oriented along the beam axis, the x axis on the horizontal plane, and the y axis oriented upward. The x and y axes define the transverse plane, while the z axis identifies the longitudinal direction. The azimuth angle ϕ is computed from the x axis. The polar angle θ is used to compute the pseudorapidity $\eta = -\log(\tan(\theta/2))$. We fix units such that $c = \hbar = 1$.

JHEP05(2019)036

Standard Model processes					
Process	Acceptance	L1 trigger efficiency	Cross section [nb]	Event fraction	Events /month
W	55.6%	68%	58	59.2%	110M
QCD	0.08%	9.6%	$1.6 \cdot 10^5$	33.8%	63M
Z	16%	77%	20	6.7%	12M
$t\bar{t}$	37%	49%	0.7	0.3%	0.6M

BSM benchmark processes				
Process	Acceptance	L1 trigger efficiency	Total efficiency	Cross-section 100 BSM events/month
$A \to 4\ell$	5%	98%	5%	0.44 pb
$LQ \to b\tau$	19%	62%	12%	0.17 pb
$h^0 \to \tau\tau$	9%	70%	6%	0.34 pb
$h^\pm \to \tau\nu$	18%	69%	12%	0.16 pb

Table 1. Acceptance and L1 trigger (i.e. p_T^ℓ and Iso requirement) efficiency for the four studied SM processes and corresponding values for the BSM benchmark models. For SM processes, we quote the total cross section before the trigger, the expected number of events per month and the fraction in the SM cocktail. For BSM models, we compute the production cross section corresponding to an average of 100 BSM events per month passing the acceptance and L1 trigger requirements. The monthly event yield is computed assuming an average luminosity per month of 5 fb^{-1}, corresponding to the running conditions discussed in section 1.

The 21 considered HLF quantities are:

- The absolute value of the isolated-lepton transverse momentum p_T^ℓ.
- The three isolation quantities (ChPFIso, NeuPFIso, GammaPFIso) for the isolated lepton, computed with respect to charged particles, neutral hadrons and photons, respectively.
- The lepton charge.
- A Boolean flag (isEle) set to 1 when the trigger lepton is an electron, 0 otherwise.
- S_T, i.e. the scalar sum of the p_T of all the jets, leptons, and photons in the event with $p_T > 30\,\text{GeV}$ and $|\eta| < 2.6$. Jets are clustered from the reconstructed PF candidates, using the `FASTJET` [33] implementation of the anti-k_T jet algorithm [34], with a jet-size parameter R=0.4.
- The number of jets entering the S_T sum (N_J).
- The invariant mass of the set of jets entering the S_T sum (M_J).
- The number of these jets being identified as originating from a b quark (N_b).

JHEP05(2019)036

- The missing transverse momentum, decomposed into its parallel ($p_{T,\parallel}^{\text{miss}}$) and orthogonal ($p_{T,\perp}^{\text{miss}}$) components with respect to the lepton ℓ direction. The missing transverse momentum is defined as the negative sum of the PF-candidate p_T vectors:

$$\vec{p}_T^{\,\text{miss}} = -\sum_q \vec{p}_T^{\,q} \,. \tag{3.2}$$

- The transverse mass, M_T, of the isolated lepton ℓ and the $\vec{p}_T^{\,\text{miss}}$ system, defined as:

$$M_T = \sqrt{2 p_T^\ell E_T^{\text{miss}} (1 - \cos\Delta\phi)}\,, \tag{3.3}$$

  with $\Delta\phi$ the azimuth separation between the $\vec{p}_T^{\,\ell}$ and $\vec{p}_T^{\,\text{miss}}$ vectors, and E_T^{miss} the magnitude of $\vec{p}_T^{\,\text{miss}}$.

- The number of selected muons (N_μ).
- The invariant mass of this set of muons (M_μ).
- The absolute value of the total transverse momentum of these muons ($p_{T,\text{TOT}}^{\mu}$).
- The number of selected electrons (N_e).
- The invariant mass of this set of electrons (M_e).
- The absolute value of the total transverse momentum of these electrons ($p_{T,\text{TOT}}^{e}$).
- The number of reconstructed charged hadrons.
- The number of reconstructed neutral hadrons.

This list of HLF quantities is not defined having in mind a specific BSM scenario. Instead, it is conceived to include relevant information to discriminate the various SM processes populating the single-lepton data stream. On the other hand, it is generic enough to allow (at least in principle) the identification of a large set of new physics scenarios.

In addition to the four SM processes listed above, we consider the following BSM models to benchmark anomaly-detection capabilities:

- A leptoquark LQ with mass 80 GeV, decaying to a b quark and a τ lepton.
- A neutral scalar boson with mass 50 GeV, decaying to two off-shell Z bosons, each forced to decay to two leptons: $A \to 4\ell$.
- A scalar boson with mass 60 GeV, decaying to two tau leptons: $h^0 \to \tau\tau$.
- A charged scalar boson with mass 60 GeV, decaying to a tau lepton and a neutrino: $h^\pm \to \tau\nu$.

For each BSM scenario, we consider any direct production mechanism implemented in `PYTHIA8`, including associate jet production. We list in table 1 the leading-order production cross section and selection efficiency for each model.

Figures 1 and 2 show the distribution of HLF quantities for the SM processes and the BSM benchmark models, respectively.

JHEP05(2019)036

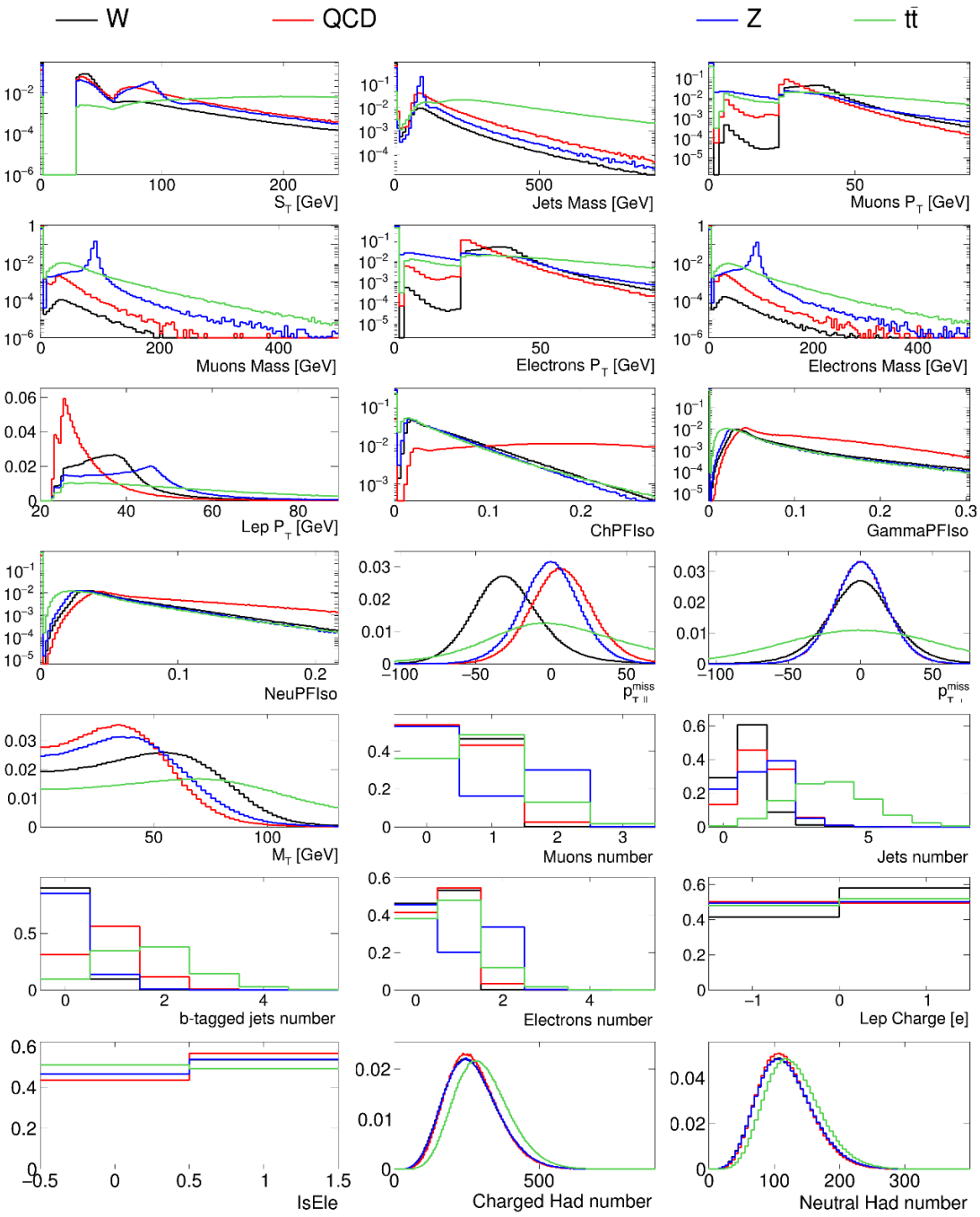

Figure 1. Distribution of the HLF quantities for the four considered SM processes.

JHEP05(2019)036

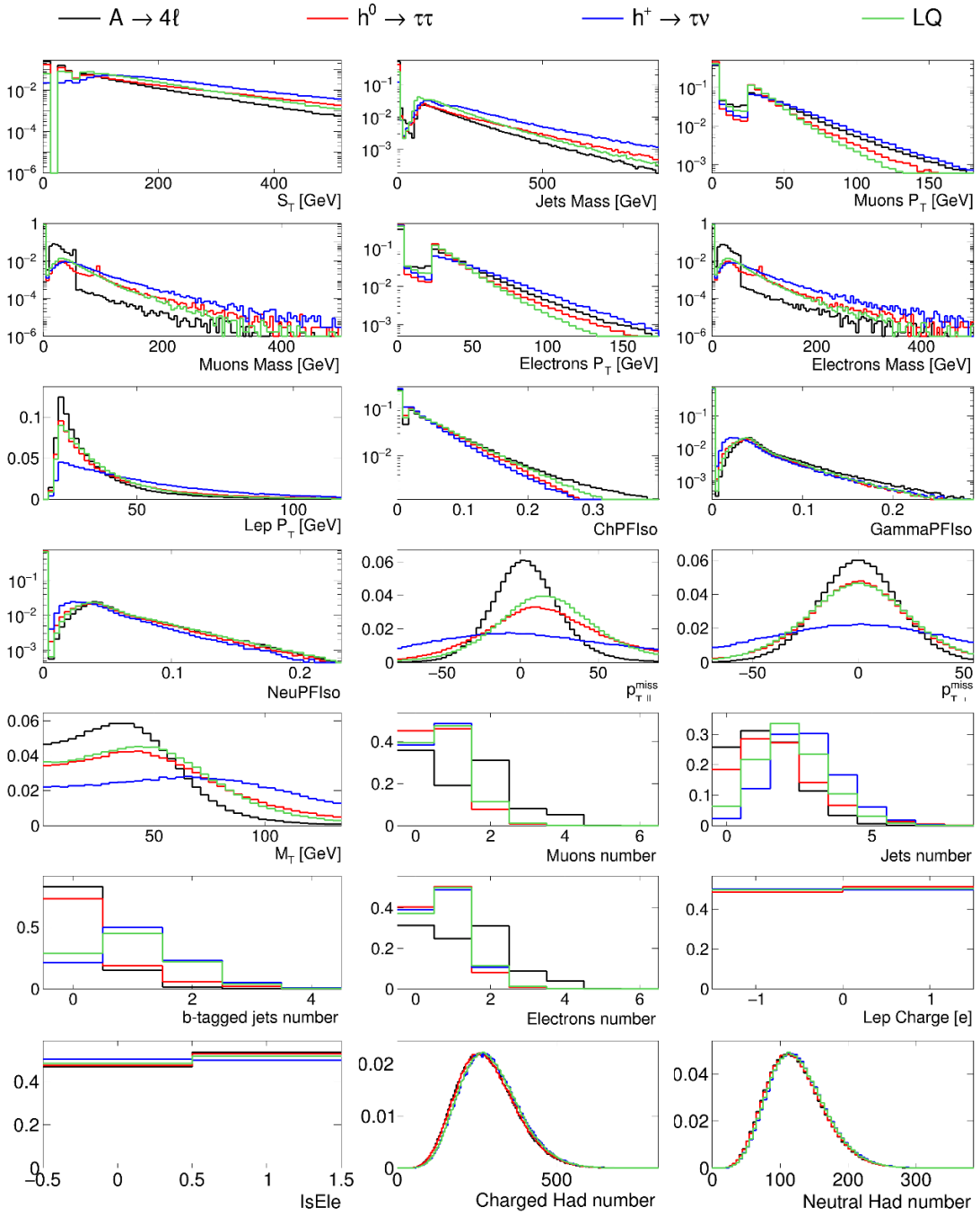

Figure 2. Distribution of the HLF quantities for the four considered BSM benchmark models.

JHEP05(2019)036

4 Model description

We train VAEs on the SM cocktail sample described in section 3, taking as input the 21 HLF quantities listed there. The use of HLF quantities to represent events limits the model independence of the anomaly detection procedure. While the list of features is chosen to represent the main physics aspects of the considered SM processes and is in no way tailored to specific BSM models, it is true that such a list might be more suitable for certain models than for others. In this respect, one cannot guarantee that the anomaly-detection performance observed on a given BSM model would generalize to any BSM scenario. We will address in a future work a possible solution to reduce the residual model dependence implied by the input event representation.

In this section, we present both the best-performing autoencoder model, trained to encode and decode the SM training sample, and a set of four supervised classifiers, each trained to distinguish one of the four BSM benchmark models from SM events. We use the classification performance of these supervised algorithms as an estimate of the best performance that the VAE could get to.

4.1 Autoencoders

Autoencoders are algorithms that compress a given set of inputs variables in a latent space (encoding) and then, starting from the latent space, reconstruct the HLF input values (decoding). The loss distribution of an AE is used in the context of anomaly detection to isolate potential anomalies. Since the compression capability learned on a given sample doesn't typically generalize to other samples, the tails of the loss distribution could be enriched by new kinds of events, different than those used to train the model. In the specific case considered in this study, the tail of the loss distribution for an AE trained on SM data might be enriched with BSM events.

In this work we focus on VAEs [11]. For each event, a plain AE predicts an encoded point in the latent space and a decoded point in the original space. In other words, AEs are point-estimate algorithms. VAEs, instead, associate to each input event an estimated probability distributions in the latent space and in the original space. Doing so, VAEs provide both a best-point estimate and an estimate of the associated statistical noise. Besides this conceptual difference, VAEs have been shown to provide competitive performances for novelty [29] and anomaly [12] detection.

We consider the VAE architecture shown in figure 3, characterized by a four-dimensional latent space. Each latent dimension is associated to a Gaussian pdf and its two degrees of freedom (mean μ_z and RMS σ_z). The input layer consists of 21 nodes, corresponding to the 21 HLF quantities described in section 3. This layer is connected to the latent space through a stack of two fully connected layers, each consisting of 50 nodes with ReLU activation functions. Two four-node layers are fully connected to the second 50-node layer. Linear activation functions are used for the first of these four-node layers, interpreted as the set of four μ_z of the four-dimension Gaussian pdf $p(z)$. The nodes of the

JHEP05(2019)036

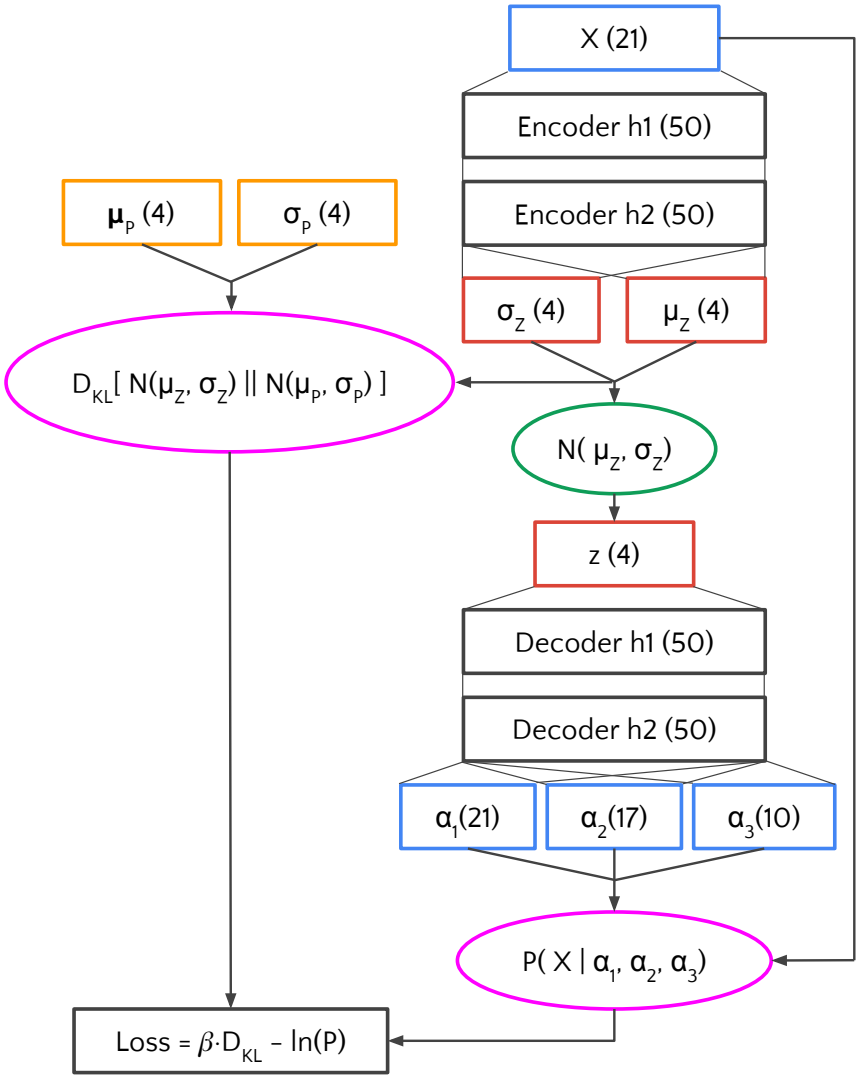

Figure 3. Schematic representation of the VAE architecture presented in the text. The size of each layer is indicated by the value within brackets. The blue rectangle X represents the input layer, which is connected to a stack of two consecutive fully connected layers (black boxes). The last of the two black box is connected to two layers with four nodes each (red boxes), representing the μ_z and σ_z parameters of the encoder pdf $p(z|x)$. The green oval represents the sampling operator, which returns a set of values for the 4-dimensional latent variables z. These values are fed into the decoder, consisting of two consecutive hidden layers of 50 nodes each (black boxes). The last of the decoder hidden layer is connected to the three output layers, whose nodes correspond to the parameters of the predicted distribution in the initial 21-dimension space. The pink ovals represent the computation of the two parts of the loss function: the KL loss and the reconstruction loss (see text). The computation of the KL requires 8 additional learnable parameters (μ_p and σ_p, represented by the orange boxes on the top-left part of the figure), corresponding to the means and RMS of the four-dimensional Gaussian prior $p(z)$. The total loss in computed as described by the formula in the bottom-left black box (see eq. (4.3)).

JHEP05(2019)036

second layer are activated by the functions:

$$\text{p-ISRLu}(x) = 1 + 5 \cdot 10^{-3} + \Theta(x)x + \Theta(-x)\frac{x}{\sqrt{1+x^2}} \ . \tag{4.1}$$

This activation allows to improve the training stability, being strictly positive defined, non linear, and with no exponentially growing term (which might have created instabilities in the early epochs of the training). The four nodes of this layer are interpreted as the σ_z parameters of $p(z)$. After several trials, the dimension of the latent space has been set to 4 in order to keep a good training stability without impacting the VAE performances. The decoding step originates from a point in the latent space, sampled according to the predicted pdf (green oval in figure 3). The coordinates of this point in the latent space are fed into a sequence of two hidden dense layers, each consisting of 50 neurons with ReLU activation functions. The last of these layers is connected to three dense layers of 21, 17, and 10 neurons, activated by linear, p-ISRLu and clipped-tanh functions, respectively. The clipped-tanh function if written as:

$$C_{\tanh}(x) = \frac{1}{2}(1 + 0.999 \cdot \tanh x) \ . \tag{4.2}$$

Given the latent-space representation, the 48 output nodes represent the parameters of the pdfs describing the input HLF probability, i.e., the α parameters of eq. (4.5).

The total VAE loss function Loss_{Tot} is a weighted sum of two pieces [35]: a term related to the reconstruction likelihood ($\text{Loss}_{\text{reco}}$) and the Kullback-Leibler divergence (D_{KL}) between the latent space pdf and the prior:

$$\text{Loss}_{\text{Tot}} = \text{Loss}_{\text{reco}} + \beta D_{\text{KL}} \, , \tag{4.3}$$

where β is a free parameter. We fix $\beta = 0.3$, for which we obtained good reconstruction performances.[4] The prior $p(z)$ chosen for the latent space is a four-dimension Gaussian with a diagonal covariance matrix. The means (μ_P) and the diagonal terms of the covariance matrix (σ_P) are free parameters of the algorithm and are optimized during the back-propagation. The Kullback-Leibler divergence between two Gaussian distributions has an analytic form. Hence, for each batch, D_{KL} can be expressed as:

$$\begin{aligned} D_{\text{KL}} &= \frac{1}{k}\sum_i D_{\text{KL}}\left(N(\mu_z^i, \sigma_z^i) \ || \ N(\mu_P, \sigma_P)\right) \\ &= \frac{1}{2k}\sum_{i,j}\left(\sigma_P^j \sigma_z^{i,j}\right)^2 + \left(\frac{\mu_P^j - \mu_z^{i,j}}{\sigma_P^j}\right)^2 + \ln\frac{\sigma_P^j}{\sigma_z^{i,j}} - 1 \, , \end{aligned} \tag{4.4}$$

where k is the batch size, i runs over the samples and j over the latent space dimensions. Similarly, $\text{Loss}_{\text{reco}}$ is the average negative-log-likelihood of the inputs given the predicted α values:

$$\begin{aligned} \text{Loss}_{\text{reco}} &= -\frac{1}{k}\sum_i \ln\left[P(x \mid \alpha_1, \alpha_2, \alpha_3)\right] \\ &= -\frac{1}{k}\sum_{i,j} \ln\left[f_j(x_{i,j} \mid \alpha_1^{i,j}, \alpha_2^{i,j}, \alpha_3^{i,j})\right] \ . \end{aligned} \tag{4.5}$$

[4] Following ref. [35], we tried to increase the value of β up to 4 without observing a substantial difference in performance.

JHEP05(2019)036

In the equation, j runs over the input space dimensions, f_j is the functional form chose to describe the pdf of the j-th input variable and $\alpha_m^{i,j}$ are the parameter of the function. Different functional forms have been chosen for f_j, to properly describe different classes of HLF distributions:

- *Clipped Log-normal + δ function*: used to describe S_T, M_J, p_T^{μ}, M_{μ}, p_T^{e}, M_e, p_T^{ℓ}, ChPFIso, NeuPFIso and GammaPFIso:

$$P(x \mid \alpha_1, \alpha_2, \alpha_3) = \begin{cases} \alpha_3 \delta(x) + \frac{1-\alpha_3}{x\alpha_2\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \alpha_1)^2}{2\alpha_2^2}\right) & \text{for } x \geq 10^{-4} \\ 0 & \text{for } x < 10^{-4} \end{cases} . \tag{4.6}$$

- *Gaussian*: used for $p_{T,\parallel}^{\text{miss}}$ and $p_{T,\perp}^{\text{miss}}$:

$$P(x \mid \alpha_1, \alpha_2) = \frac{1}{\alpha_2\sqrt{2\pi}} \exp\left(-\frac{(x-\alpha_1)^2}{2\alpha_2^2}\right) . \tag{4.7}$$

- *Truncated Gaussian*: a Gaussian function truncated for negative values and normalized to unit area for $X > 0$. Used to model M_T:

$$P(x \mid \alpha_1, \alpha_2) = \Theta(x) \cdot \frac{1 + 0.5 \cdot (1 + \text{erf}\frac{-\alpha_1}{\alpha_2\sqrt{2}})}{\alpha_2\sqrt{2\pi}} \exp\left(-\frac{(x-\alpha_1)^2}{2\alpha_2^2}\right) . \tag{4.8}$$

- *Discrete truncated Gaussian*: like the truncated Gaussian, but normalized to be evaluated on integers (i.e. $\sum_{n=0}^{\infty} P(n) = 1$). This function is used to describe N_{μ}, N_e, N_b and N_J. It is written as:

$$P(n \mid \alpha_1, \alpha_2) = \Theta(x) \left[\text{erf}\left(\frac{n + 0.5 - \alpha_1}{\alpha_2\sqrt{2}}\right) - \text{erf}\left(\frac{n - 0.5 - \alpha_1}{\alpha_2\sqrt{2}}\right)\right] \mathcal{N} , \tag{4.9}$$

  where the normalization factor $\mathcal{N}$ is set to:

$$\mathcal{N} = 1 + \frac{1}{2}\left(1 + \text{erf}\left(\frac{-0.5 - \alpha_1}{\alpha_2\sqrt{2}}\right)\right) . \tag{4.10}$$

- *Binomial*: used for (isEle) and lepton charge:

$$P(n \mid p) = \delta_{n,m} p + \delta_{n,l}(1 - p) , \tag{4.11}$$

  where m and l are the two possible values of the variable (0 or 1 for (isEle) and -1 or 1 for lepton charge) and $p = C_{\tanh}(\alpha_1)$

- *Poisson*: used for charged-particle and neutral-hadron multiplicities:

$$P(n \mid \mu) = \frac{\mu^n e^{-\mu}}{\Gamma(n+1)} , \tag{4.12}$$

  where $\mu = \text{p-ISRLu}(\alpha_1)$.

JHEP05(2019)036

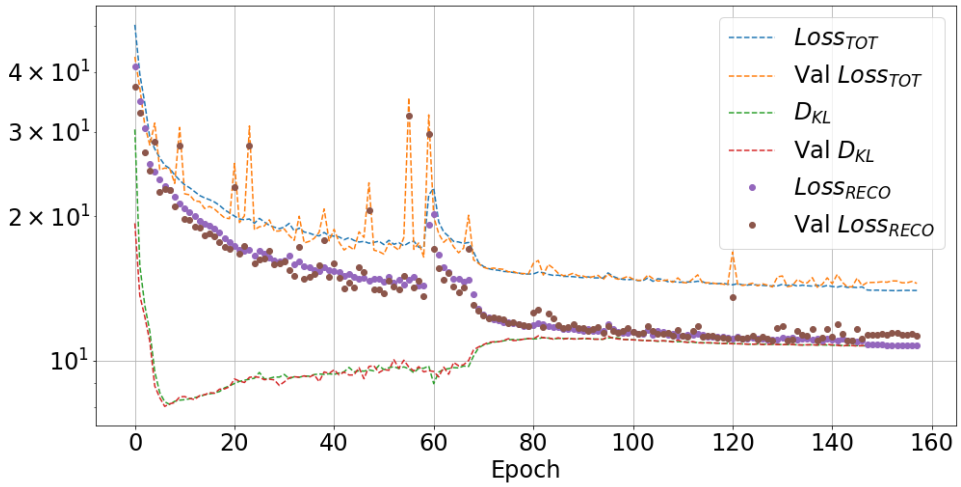

Figure 4. Training history for VAE. Total loss, reconstruction negative-log-likelihood ($\text{Loss}_{\text{reoc}}$) and KL divergence ($D_{\text{KL}}$) are shown separately for training and validation set though all the training epochs.

These custom functions provide an improved performance with respect to the standard choice of an MSE loss. When using the MSE loss, one is implicitly writing the likelihood of the input quantities as a product of Gaussian functions with equal variance. This choice is clearly a poor description of the input distributions at hand in this application and it results in a poor representation of the cores and the tails of the input distributions. Instead, the use of these tailored functions allows to correctly describe the distribution cores and to improve the description of the tails.

We point out that the final performance depends on the choice of the p(x—z) functional form (i.e., on the modeled dependence of the observed features on the latent variables) and the p(z) prior function. The former was tuned looking at the distributions for SM events. The latter is arbitrary. We explored techniques to optimize the choice of p(z), learning it from the data [36]. In this case, no practical advantage in terms of anomaly detection was observed. An improved choice of p(x—z) and the possibility of learning p(z) during the train could potentially further boost the performances of this algorithm and will be the subject of future studies with real LHC collision data.

The model shown in figure 3 is implemented in `KERAS+TENSORFLOW` [37, 38], trained with the Adam optimizer [39] on a SM dataset of 3.45M events, equivalent to an integrated luminosity of $\sim$ 100 pb^{-1}. The SM validation dataset is made of 3.45M of statistically independent examples. Such a sample would be collected in about ten hours of continuous run, under the assumptions made in this study (see section 1). In training, we fix the batch size to 1000. We use early stopping with patience set to 20 and $\delta_{\text{min}} = 0.005$, and we progressively reduce the learning rate on plateau, with patience set to 8 and $\delta_{\text{min}} = 0.01$.

The model's training history is shown in figure 4. Figure 5 shows the comparison of the input and output distributions for the 21 HLF quantities in the validation dataset. A general good agreement is observed on the bulk of the distributions, even if some of

JHEP05(2019)036

Process	AUC	TPR [%]
$A \to 4\ell$	0.98	5.4
$LQ \to b\tau$	0.94	0.2
$h^0 \to \tau\tau$	0.90	0.1
$h^\pm \to \tau\nu$	0.97	0.3

Table 2. Classification performance of the four BDT classifiers described in the text, each trained on one of the four BSM benchmark models. The two set of values correspond to the area under ROC curve (AUC), and to the true positive rate (TPR) for a SM false positive rate $\epsilon_{\rm SM} = 5.4 \cdot 10^{-6}$, i.e., to ~ 1000 SM events accepted every month.

the distributions are not well described on the tails. These discrepancies don't have a sizable impact on the anomaly-detection strategy, as shown in section. 5. Nevertheless, alternative architectures were tested, in order to reduce these discrepancies. For instance, we increased or decreased the dimensionality of the latent space, we changed the value of β in eq. (4.3), we changed the number of neurons in the hidden layers, tried the RMSprop optimizer, and used plain Gaussian functions to describe the 21 input features. Some of these choices improved the encoding-decoding capability of the VAE, with up to a 10% decrease of the loss function at the end of the training. On the other hand, none of these alternative models provided a sizable improvement in the anomaly-detection performance. For simplicity, we decided to limit our study to the architecture in figure 3 and dropped these alternative models.

4.2 Supervised classifiers

For each of the four BSM benchmark models, we train a fully-supervised classifier, based on a Boosted Decision Tree (BDT). Each BDT receives as input the same 21 features used by the VAE and is trained on a labeled dataset consisting of the SM cocktail (the background) and one of the four BSM benchmark models (the signal). The implementation is done through the Gradient Boosted Classifier of the scikit-learn library [40]. The algorithm was tuned with up to 150 estimators, minimum samples per leaf and maximum depth equal to 3, a learning rate of 0.1, and a tolerance of 10^{-4} on the validation loss function (choose to be the default deviance). Each BDT, tailored to a specific BSM model, is trained on 3.45M SM events and about 0.5M BSM events, consistently up-weighted in order to match the size of the SM sample during the training.

We show in table 2 and in figure 6 the classification performance of the four supervised BDTs, which set a qualitative upper limit for VAE's results. Overall, the four models can be discriminated with good accuracy, with some loss of performance for those models sharing similarities with specific SM processes (e.g., $h^0 \to \tau\tau$ exhibiting single- and double-lepton topology with missing transverse energy, typical of $t\bar{t}$ events). In the table, we also quote the true-positive rate (TPR) for each BSM model corresponding to a working point of SM false positive rate $\epsilon_{\rm SM} = 5.4 \cdot 10^{-6}$, corresponding to an average of ~ 1000 SM events accepted every month.

In addition to BDTs, we experimented with fully-connected deep neural networks (DNNs) with two hidden layers. Despite trying different architectures, we didn't find a

JHEP05(2019)036

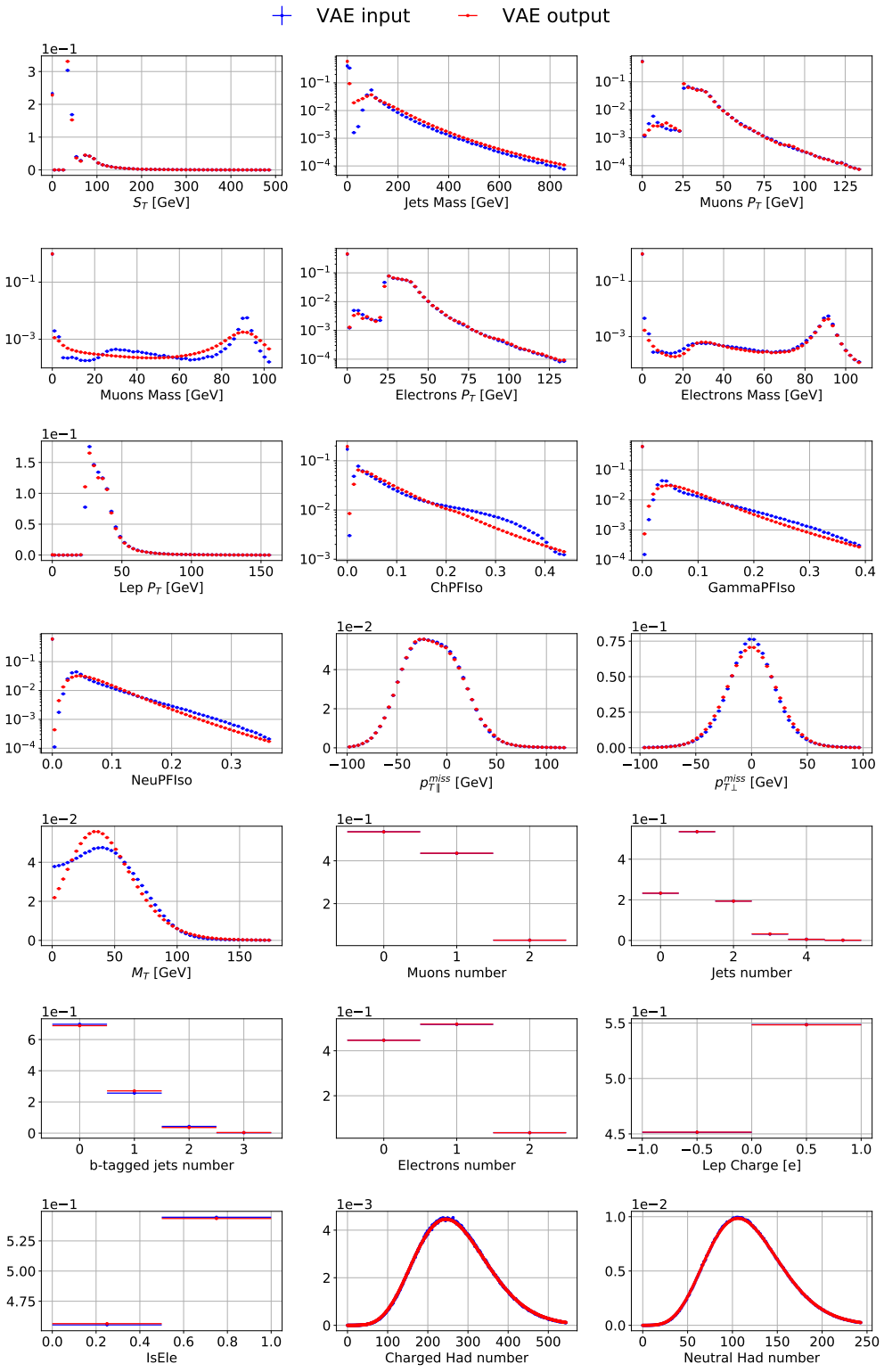

Figure 5. Comparison of input (blue) and output (red) probability distributions for the HLF quantities in the validation sample. The input distributions are normalized to unity. The output distributions are obtained summing over the predicted pdf of each event, normalized to the inverse of the total number of events (so that the total sum is normalized to unity).

JHEP05(2019)036

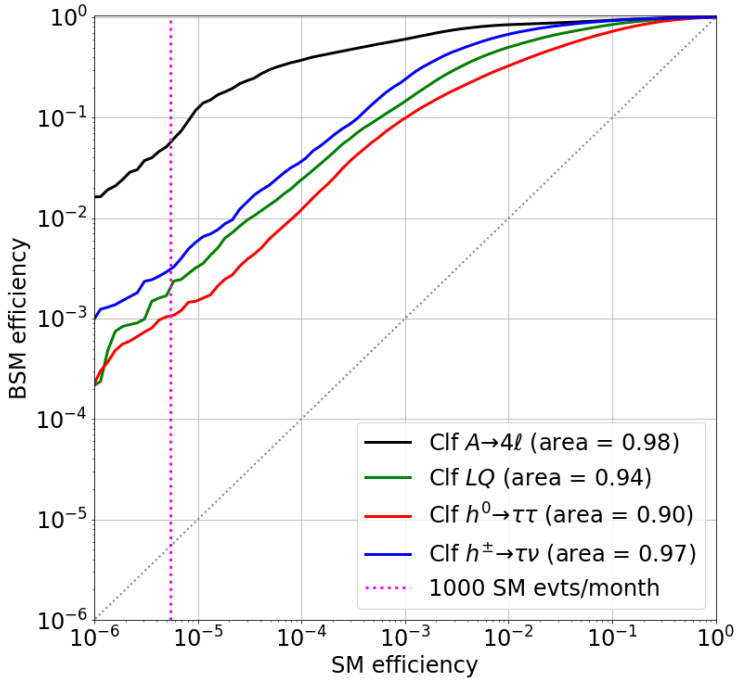

Figure 6. ROC curves for the fully-supervised BDT classifiers, optimized to separate each of the four BSM benchmark models from the SM cocktail dataset.

configuration in which the DNN classifiers could outperform the BDTs. This is due to the fact that, given the limited complexity of the problem at hand, a simple BDT can extract the maximum discrimination power from the 21 inputs. The limiting factor preventing to reach larger auc values is not to be found in the model complexity but in the discriminating power of the 21 input features. Not being tailored on the benchmark BSM scenarios, these features don't carry all the needed information for an optimal signal-to-background separation. While certainly one could obtain a better performance with more tailored classifiers, the purpose of this exercise was to provide a fair comparison for the VAE. In view of these considerations, we decided to use the BDTs as reference supervised classifiers.

5 Results with VAE

An event is classified as anomalous whenever the associated loss, computed from the VAE output, is above a given threshold. Since no BSM signal has been observed by LHC experiments so far, it is reasonable to expect that a new-physics signal, if any, would be characterized by a low production cross section and/or features very similar to those of a SM process. In view of this, we decided to use a tight threshold value, in order to reduce as much as possible any SM contribution.

Figure 7 shows the distribution of the $\text{Loss}_{\text{reco}}$ and D_{KL} loss components for the validation dataset. In both plots, the vertical line represents a lower threshold such that a fraction $\epsilon_{\text{SM}} = 5.4 \cdot 10^{-6}$ of the SM events would be retained. This threshold value would

JHEP05(2019)036

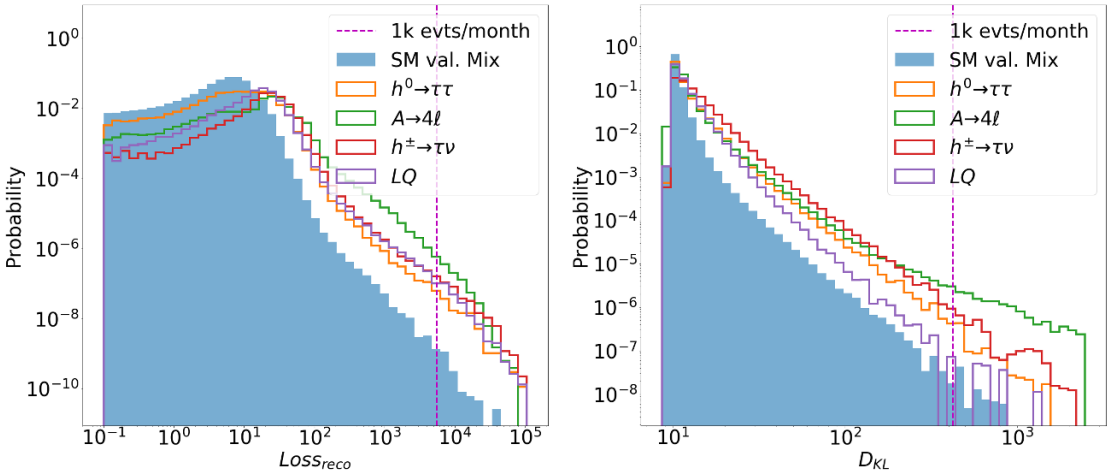

Figure 7. Distribution of the VAE's loss components, $\text{Loss}_{\text{reco}}$ (left) and D_{KL} (right), for the validation dataset. For comparison, the corresponding distribution for the four benchmark BSM models are shown. The vertical line represents a lower threshold such that $5.4 \cdot 10^{-6}$ of the SM events would be retained, equivalent to ~ 1000 expected SM events per month.

result in ~ 1000 SM events to be selected every month, i.e., a daily rate of ~ 33 SM events, as illustrated in table 3. The acceptance rate is calculated assuming the LHC running conditions listed in section 1. Table 3 also reports the by-process VAE selection efficiency and the relative background composition of the selected sample.

Figure 7 also shows the $\text{Loss}_{\text{reco}}$ and D_{KL} distributions for the four benchmark BSM models. We observe that the discrimination power, loosely quantified by the integral of these distributions above threshold, is better for $\text{Loss}_{\text{reco}}$ than D_{KL} and that the impact of the D_{KL} term on Loss_{Tot} is negligible. Anomalies are then defined as events laying on the right tail of the expected $\text{Loss}_{\text{reco}}$ distribution. Due to limited statistics in the training sample, the p-value corresponding to the chosen threshold value could be uncalibrated. This could result in a deviation of the observed rate from the expected value, an issue that one can address tuning the threshold. On the other hand, an uncalibrated p-value would also impact the number of collected BSM events, and the time needed to collect an appreciable amount of these events.

Once the $\text{Loss}_{\text{reco}}$ selection is applied, the anomalous events don't cluster on the tails of the distributions of the input features. Instead, they tend to cover the full feature-definition range. This is an indication of the fact that the VAE does more than a simple selection of feature outliers, which is what is done by traditional single-lepton trigger or by dedicated cross triggers (e.g., triggers that select events with soft leptons and large missing transverse energy, S_T, etc.). This is shown in figure 8 for SM events. A similar conclusion can be obtained from figure 9, showing the distribution of the 21 input HLF quantities for the $A \to 4\ell$ benchmark model, before and after applying the threshold requirement on the VAE loss.

The left plot in figure 10 shows the ROC curves obtained from the $\text{Loss}_{\text{reco}}$ distribution of the four BSM benchmark models and the SM cocktail, compared to the corresponding BDT curves of section 4.2. As expected, the results obtained with the supervised BDTs

JHEP05(2019)036

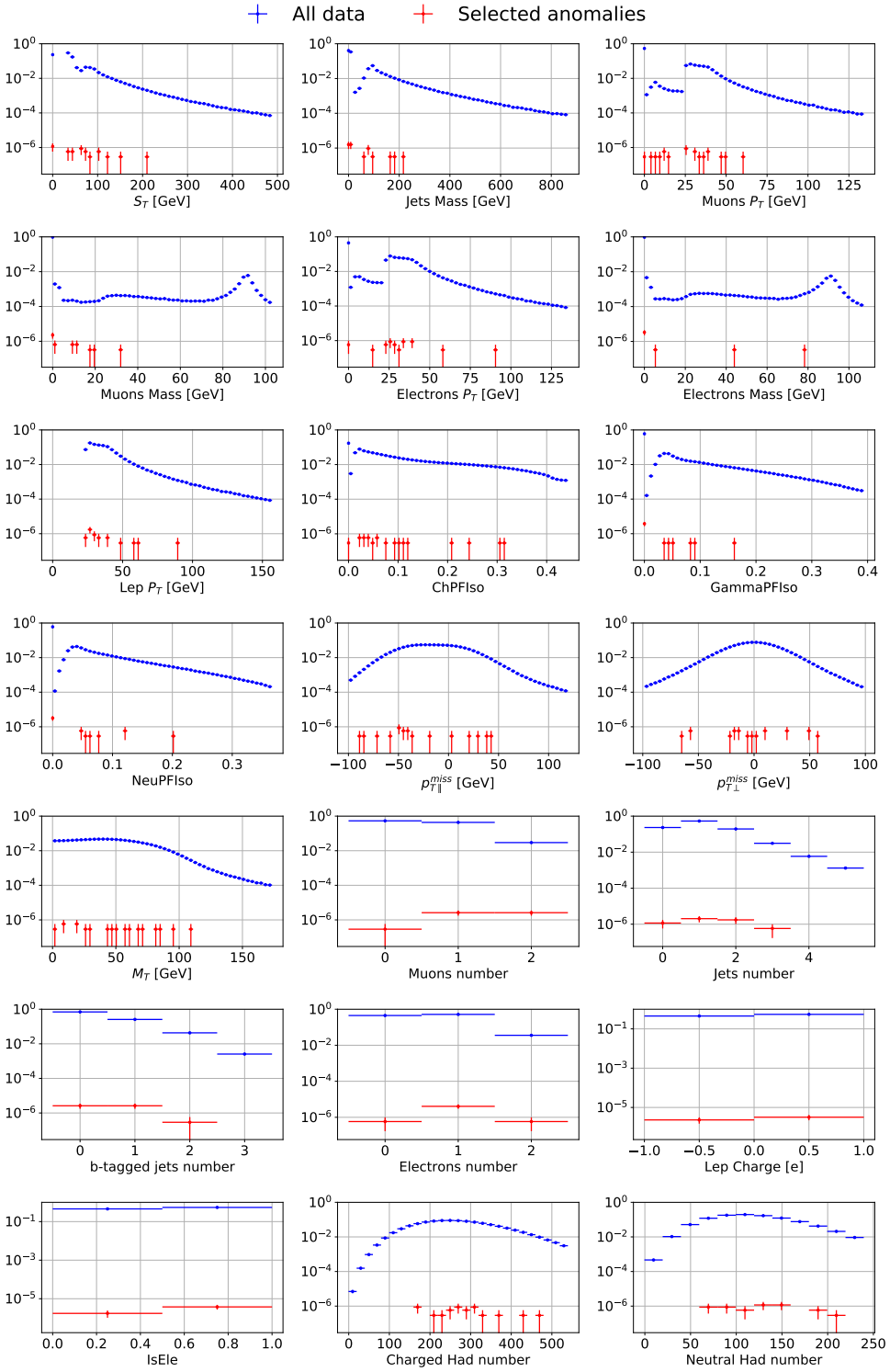

Figure 8. Comparison between the input distribution for the 21 HLF of the validation dataset (blue histograms) and the distribution of the SM outlier events selected from the same sample by applying the $\text{Loss}_{\text{reco}}$ threshold (red dots). The outlier events cover a large portion of the HLF definition range and don't cluster on the tails.

JHEP05(2019)036

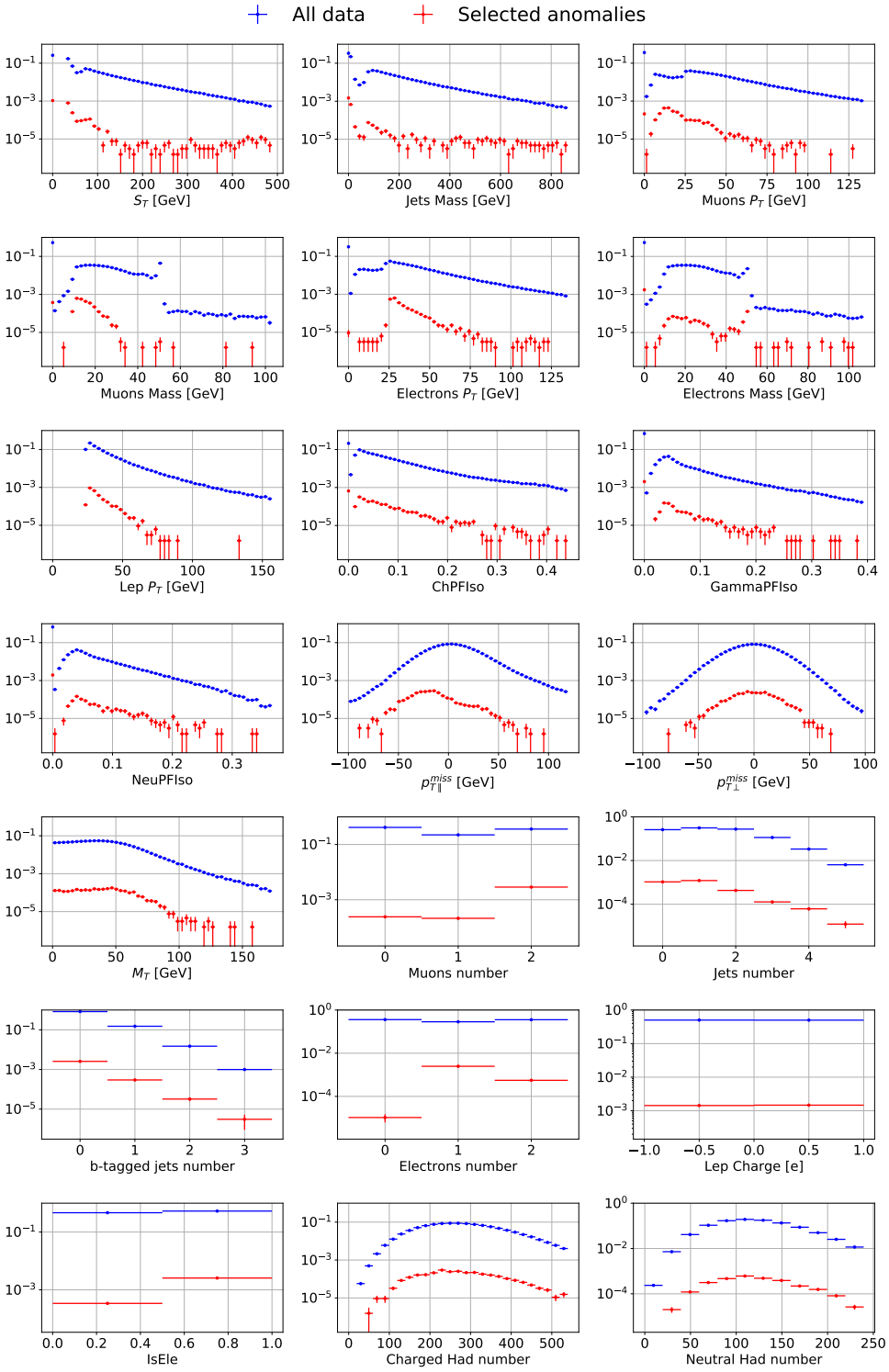

Figure 9. Comparison between the distribution of the 21 HLF distribution for $A \to 4\ell$ full dataset (blue) and $A \to 4\ell$ events selected by applying the $\text{Loss}_{\text{reco}}$ threshold (red). The selected events are not trivially sampled from the tail.

JHEP05(2019)036

Standard Model processes			
Process	VAE selection	Sample composition	Events/month
W	$3.6 \pm 0.7 \cdot 10^{-6}$	32%	379 ± 74
QCD	$6.0 \pm 2.3 \cdot 10^{-6}$	29%	357 ± 143
Z	$21 \pm 3.5 \cdot 10^{-6}$	21%	256 ± 43
$t\bar{t}$	$400 \pm 9 \cdot 10^{-6}$	18%	212 ± 5
Tot			1204 ± 167

Table 3. By-process acceptance rate for the anomaly detection algorithm described in the text, computed applying the threshold on $\text{Loss}_{\text{reco}}$ shown in figure 7. The threshold is tuned such that a fraction of about $\epsilon_{\text{SM}} = 5.4 \cdot 10^{-6}$ of SM events would be accepted, corresponding to ~ 1000 SM events/month, assuming the LHC running conditions listed in section 1. The sample composition refers to the subset of SM events accepted by the anomaly detection algorithm. All quoted uncertainties refer to 95% CL regions.

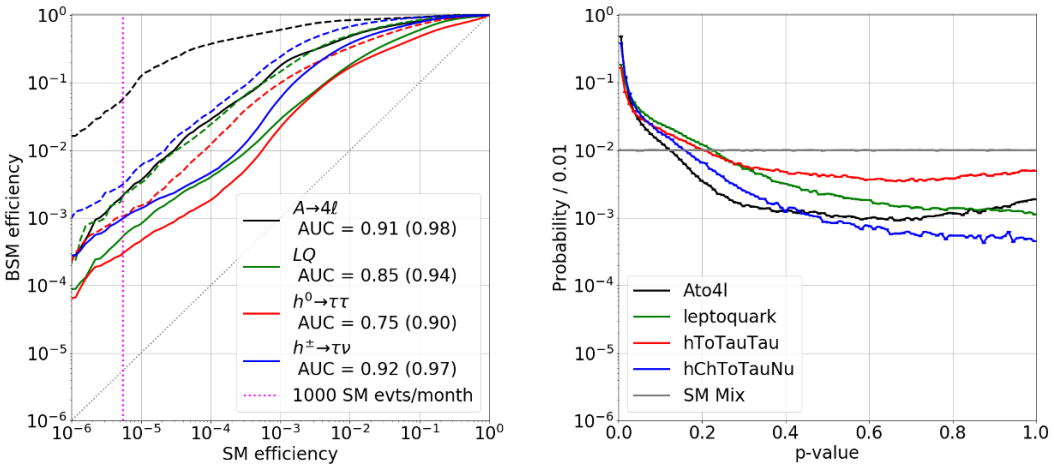

Figure 10. Left: ROC curves for the VAE trained only on SM events (solid), compared to the corresponding curves for the four supervised BDT models (dashed) described in section 4.2. Right: normalized p-value distribution distribution for the SM cocktail events and the four BSM benchmark processes.

outperform the VAE. On the other hand, the VAE can probe at the same time the four scenarios with comparable performances. This is a consequence of the trade off between precision and model independence and an illustration of the complementarity between the approach presented in this work and traditional supervised techniques. The right plot in figure 10 shows the one-sided p-value computed from the cocktail SM distribution, both for the SM events themselves (flat by construction) and for the four BSM processes. As the plot shows, BSM processes tend to concentrate at small p-values, which allows their identification as anomalies.

Table 4 summarizes the VAE's performance on the four BSM benchmark models. Together with the selection efficiency corresponding to $\epsilon_{\text{SM}} = 5.4 \cdot 10^{-6}$, the table reports the effective cross section (cross section after applying the trigger requirements) that would

JHEP05(2019)036

BSM benchmark processes			
Process	VAE selection efficiency	Cross-section 100 events/month [pb]	Cross-section S/B = 1/3 [pb]
$A \to 4\ell$	$2.8 \cdot 10^{-3}$	7.1	27
$LQ \to b\tau$	$6.7 \cdot 10^{-4}$	30	110
$h^0 \to \tau\tau$	$3.6 \cdot 10^{-4}$	55	210
$h^{\pm} \to \tau\nu$	$1.2 \cdot 10^{-3}$	17	65

Table 4. Breakdown of BSM processes efficiency, and cross section values corresponding to 100 selected events in a month and to a signal-over-background ratio of 1/3 (i.e., an absolute yield of $\sim$ 400 events/month). The monthly event yield is computed assuming an average luminosity per month of 5 fb^{-1}, computing by taking the LHC 2016 data delivery ($\sim$ 40 fb^{-1} collected in 8 months). All quoted efficiencies are computed fixing the VAE loss threshold $\epsilon_{\text{SM}} = 5.4 \cdot 10^{-6}$.

correspond to 100 BSM events selected in a month (assuming an integrated luminosity of 5 fb^{-1}). Similarly, we quote the cross section that would result in a signal-to-background ratio of 1/3 on the sample of events selected by the VAE. The VAE can probe the four models down to small cross section values, comparable to the existing exclusion bounds for these mass ranges. As an example, ref. [41] excludes a $LQ \to \tau b$ with a mass of 150 GeV and production cross section larger than $\sim$ 10 pb, using 4.8 fb^{-1} at a center-of-mass energy of 7 TeV, while most recent searches [42] cannot cover such a low mass value, due to trigger limitations.

Unlike a traditional trigger strategy, a VAE-based selection is mainly intended to select a high-purity sample of interesting event, at the cost of a typically small selection efficiency. To demonstrate this point, we consider a sample selected with the VAE and one selected using a typical inclusive single lepton trigger (SLT), consisting on a tighter selection than the one described in section 3. In particular, we require $p_T^{\ell} > 27$ GeV and ISO < 0.25. We consider the signal-over-background ratio (SBR) for the VAE's threshold selection and the SLT. While these quantities depend on the production cross section of the considered BSM model, their ratio

$$\frac{\text{SBR}_{\text{VAE}}}{\text{SBR}_{\text{SLT}}} = \left(\frac{\epsilon_{\text{SLT}}}{\epsilon_{\text{VAE}}}\right)_{\text{SM}} \cdot \left(\frac{\epsilon_{\text{VAE}}}{\epsilon_{\text{SLT}}}\right)_{BSM} \tag{5.1}$$

is only a function of the selection efficiency for the SLT (ϵ_{SLT}) and the for the VAE ϵ_{VAE} for SM and BSM events. Table 5 shows how the SBR reached by the VAE is about two order of magnitude larger than what a traditional inclusive SLT could reach.

6 How to deploy a VAE for BSM detection

The work presented in this paper suggests the possibility of deploying a VAE as a trigger algorithms associated to dedicated data streams. This trigger would isolate anomalous events, similarly to what was done by the CMS experiment at the beginning of the first LHC run. With early new physics signal being a possibility at the LHC start, the CMS

JHEP05(2019)036

	SM	$A \to 4\ell$	$LQ \to b\tau$	$h^0 \to \tau\tau$	$h^\pm \to \tau\nu$
ϵ_{VAE}	$5.3 \cdot 10^{-6}$	$2.8 \cdot 10^{-3}$	$6.7 \cdot 10^{-4}$	$3.6 \cdot 10^{-4}$	$1.2 \cdot 10^{-3}$
ϵ_{SLT}	0.6	0.5	0.6	0.7	0.6
$\epsilon_{SLT}/\epsilon_{\text{VAE}}$	$1.1 \cdot 10^5$	$1.8 \cdot 10^2$	$9.0 \cdot 10^2$	$1.7 \cdot 10^3$	$5.8 \cdot 10^2$
$\text{SBR}_{\text{VAE}}/\text{SBR}_{\text{SLT}}$	—	625	125	70	191

Table 5. Selection efficiencies for a typical single lepton trigger (SLT) and the proposed VAE selection, shown for the four benchmark BSM models and for the SM cocktail. The last row quotes the corresponding BSM-to-SM ratio of signal-over-background ratios (SBRs), quantifying the purity of the selected sample.

experiment deployed online a set of algorithms (collectively called *hot line*) to select potentially interesting new-physics candidates. At that time, anomalies were characterized as events with high-p_T particles or high particle multiplicities, in line with the kind of early-discovery new physics scenarios considered at that time. The events populating the hot-line stream were immediately processed at the CERN computing center (as opposed to traditional physics streams, that are processed after 48 hours). The hot-line algorithms were tuned to collect O(10) events per day, which were then visually inspected by experts.

While the focus of the work presented in this paper is not an early discovery, the spirit of the application we propose would be similar: a set of VAEs deployed online would select a limited number of events every day. These events would be collected in a dedicated dataset and further analyzed. The analysis technique could go from visual inspection of the collisions to detailed studies of reconstructed objects, up to some kind of model-independent analysis of the collected dataset, e.g. a deep-learning implementation of a model-independent hypothesis testing [16] directly on the loss distribution (provided a reliable sample of background-only data).

While a pure SM sample to train VAEs could only be obtained from a MC simulation, the presence of outlier contamination in the training sample has typically a tiny impact on performance. One could then imagine to train the VAE models on so-far collected data and use them on the events entering the HLT system. Such a training could happen offline on a dedicated dataset, e.g., deploying triggers randomly selecting events entering the last stage of the trigger system. The training could even happen online, assuming the availability of sufficient computing resources. As it happens with normal triggers, at the very beginning one would use some MC sample or some control sample from previously collected data to estimate the threshold corresponding to the target SM rate. Then, as it happens normally during HLT operations, the threshold will have to be monitored on real data and adjusted if needed.

To demonstrate the feasibility of a train-on-data strategy, we enrich the dataset used in section 4 with a signal contamination of $A \to 4\ell$ events. As a starting point, the amount of injected signal is tuned to a luminosity of 100 pb^{-1} and a cross section of 7.1 pb, corresponding to the value at which the VAE in section 4 would select 100 $A \to 4\ell$ events in one month. This results into about 700 $A \to 4\ell$ events added to the training sample. The VAE is trained following the procedure outlined in section 4 and its performance is

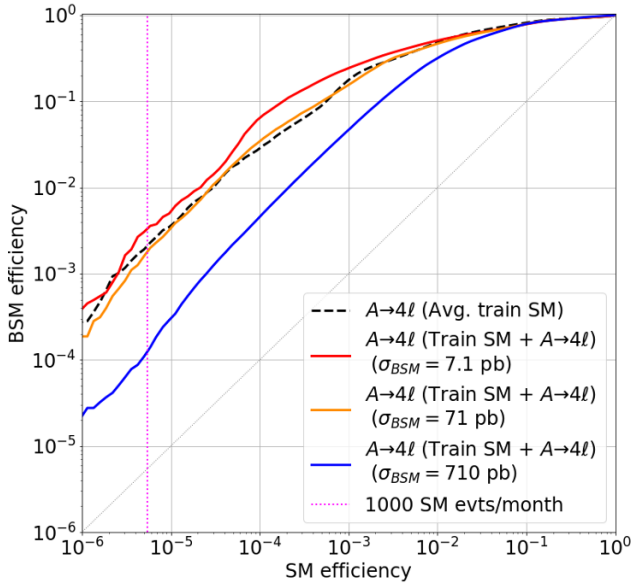

Figure 11. ROC curves for the VAE trained on SM contaminated with and without $A \to 4\mu$ contamination. Different levels of contamination are reported corresponding to 0.02% ($\sigma = 7.15$ pb — equal to the estimated one to have 100 events per month), 0.19% ($\sigma = 71.5$ pb) and 1.89% ($\sigma = 715$ pb) of the training sample.

compared to that obtained on a signal-free dataset of the same size. The comparison of the ROC curves for the two models is shown in figure 11. In the same figure, we show similar results, derived injecting a $\times 10$ and $\times 100$ signal contamination. A performance degradation is observed once the signal cross section is set to 710 pb (i.e., 100 times larger than the sensitivity value found in section 4). At that point, the contamination is so large that the signal becomes as abundant as $t\bar{t}$ events and would have easily detectable consequences. For comparison, at a production cross section of 27 pb a third of the events selected by the VAE in section 4 would come from $A \to 4\ell$ production (see table 4). Such a large yield would still have negligible consequences on the training quality. This test shows that a robust anomaly-detecting VAE could be trained directly on data, even in presence of previously undetected (e.g., at Tevatron, 7 TeV and 8-TeV LHC) BSM signals.

The possibility of training the VAE on data would substantially simplify the implementation of the strategy proposed in this work, since any possible systematic bias in the data would be automatically taken into account during the training process. In addition, it would make the procedure robust against other systematic effects (e.g., energy scale, efficiency, etc.) that would affect a MC-based training.

7 Conclusions and outlook

We present a strategy to isolate potential BSM events produced by the LHC, using variational autoencoders trained on a reference SM sample. Such an algorithm could be used in the trigger system of general-purpose LHC experiments to identify recurrent anomalies,

JHEP05(2019)036

which might otherwise escape observation (e.g., being filtered out by a typical trigger selection). Taking as an example a single-lepton data stream, we show how such an algorithm could select datasets enriched with events originating from challenging BSM models. We also discuss how the algorithm could be trained directly on data, with no sizable performance loss, more robustness against systematic uncertainties, and a big simplification of the training and deployment procedure.

The main purpose of such an application is not to enhance the signal selection efficiency for BSM models. Indeed, this application is tuned to provide a high-purity sample of potentially interesting events. We showed that events produced by not-yet-excluded BSM models with cross sections in the range of $\mathcal{O}(10)$ to $\mathcal{O}(100)$ pb could be isolated in a $\sim 30\%$ pure sample of ~ 43 events selected per day. The price to pay to reach such a purity is a relatively small signal efficiency and a strong bias in the dataset definition, which makes these events marginal and difficult to use in a traditional data-driven and supervised search for new physics.

The final outcome of this application would be a list of anomalous events, that the experimental collaborations could further scrutinize and even release as a catalog, similarly to what is typically done in other scientific domains. Repeated patterns in these events could motivate new scenarios for beyond-the-standard-model physics and inspire new searches, to be performed on future data with traditional supervised approaches.

We stress the fact that the power of the proposed approach is in its generality and not in its sensitivity to a particular BSM scenario. We show that a simple BDT could give a better discrimination capability for a given BSM hypothesis. On the other hand, such a supervised algorithm would not generalize to other BSM scenarios. The VAE, instead, comes with little model dependence and therefore generalizes to unforeseen BSM models. On the other hand, the VAE cannot guarantee an optimal performance in any scenario. As typical of autoencoders used for anomaly detection, our VAE model is trained to learn the SM background at best, but there is no guarantee that the best SM-learning model will be the best anomaly detection algorithm. By definition, the anomaly detection capability of the algorithm does not enter the loss function, as well as, by construction, no signal event enters the training sample. This is the price to pay when trading discrimination power for model independence.

We believe that such an application could help extending the physics reach of the current and next stages of the CERN LHC. The proposed strategy is demonstrated for a single-lepton data stream coming from a typical L1 selection. On the other hand, this approach could be generalized to any other data stream coming from any L1 selection, so that the full ~ 100 Hz rate entering the HLT system of ATLAS or CMS could be scrutinized. While the L1 selection still represents a potentially dangerous bias, an algorithm running in the HLT could access 100 times more events than the ~ 1 kHz stream typically available for offline studies. Moreover, thanks to progresses in the deployment of deep neural networks on FPGA boards [43], it is conceivable that VAEs for anomaly detection could be also deployed in the L1 trigger systems in a near future. In this way, the VAE would access the full L1 input data stream.

JHEP05(2019)036

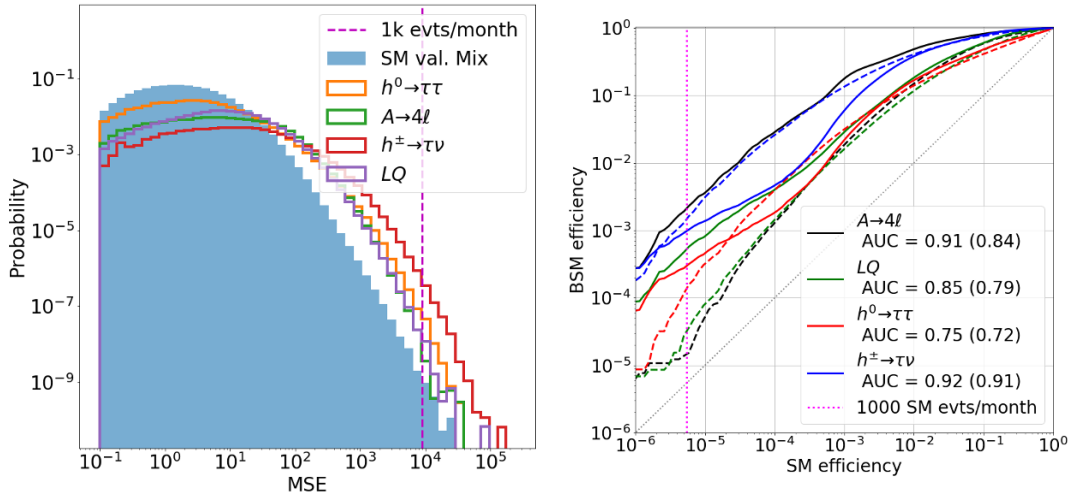

Figure 12. Left: distribution of the AE loss (MSE) for the validation dataset. The distribution for the SM processes and the four benchmark BSM models are shown. Right: ROC curves for the AE (dashed lines) trained only on SM mix, compared to the corresponding VAE curves from figure 10 (solid). The vertical dotted line represents the $\epsilon_{\text{SM}} = 5.4 \cdot 10^{-6}$ threshold considered in this study.

Acknowledgments

We thank D. Rezende for his precious suggestions, which motivated us to explore Variational Autoencoders for this work. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement n^o 772369) and the United States Department of Energy, Office of High Energy Physics Research under Caltech Contract No. DE-SC0011925. This work was conducted at "*iBanks*", the AI GPU cluster at Caltech. We acknowledge NVIDIA, SuperMicro and the Kavli Foundation for their support of "*iBanks*".

A Comparison with auto-encoder

For sake of completeness, we repeated the strategy presented in this work on a simple AE. The architecture was fixed to be as close as possible to that of the VAE introduced in section 4. The change from VAE to AE imply these two changes: the output layer has the same dimensionality of the input layer; the latent layer includes four neurons (as opposed to 8), corresponding to the four latent variables z (and not to the μ and σ parameters of the z distribution). An MSE loss function is used. The optimizer and callbacks used to trained the VAE are are used in this case. Figure 12 shows the loss function distribution and a comparison between the ROC curves of the VAE and AE. These distributions directly compare to the left plots of figures 7 and 10, since in that case only the reconstruction part of the loss was used. For convenience, the VAE ROC curves are also shown here, represented by the dashed lines. When considering the four BSM benchmark models presented in this work, the AE provides competitive performances, for some choice of the SM accepted-

JHEP05(2019)036

event rate. On the other hand, the VAE usually outperforms a plain AE for the rate considered in this study ($\epsilon_{\text{SM}} = 5.4 \cdot 10^{-6}$). With the exception of the $h^{\pm} \to \tau\nu$ model (for which the AE provides a 30% larger efficiency than the VAE), the VAE provides larger efficiency on the BSM models, with improvements as large as two orders of magnitude (for the $A \to 4\ell$ model).

References

[1] ATLAS, CMS, LHC Higgs Combination Group collaboration, *Procedure for the LHC Higgs boson search combination in summer 2011*, CMS-NOTE-2011-005 (2011).

[2] ATLAS collaboration, *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*, *Phys. Lett.* **B 716** (2012) 1 [arXiv:1207.7214] [INSPIRE].

[3] CMS collaboration, *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*, *Phys. Lett.* **B 716** (2012) 30 [arXiv:1207.7235] [INSPIRE].

[4] CDF collaboration, *Global search for new physics with* $2.0\,fb^{-1}$ *at CDF*, *Phys. Rev.* **D 79** (2009) 011101 [arXiv:0809.3781] [INSPIRE].

[5] D0 collaboration, *Model independent search for new phenomena in* $p\bar{p}$ *collisions at* $\sqrt{s} = 1.96$ *TeV*, *Phys. Rev.* **D 85** (2012) 092015 [arXiv:1108.5362] [INSPIRE].

[6] H1 collaboration, *A general search for new phenomena at HERA*, *Phys. Lett.* **B 674** (2009) 257 [arXiv:0901.0507] [INSPIRE].

[7] CMS collaboration, *MUSiC, a model unspecific search for new physics, in pp collisions at* $\sqrt{s} = 8$ *TeV*, CMS-PAS-EXO-14-016 (2256653).

[8] ATLAS collaboration, *A strategy for a general search for new phenomena using data-derived signal regions and its application within the ATLAS experiment*, *Eur. Phys. J.* **C 79** (2019) 120 [arXiv:1807.07447] [INSPIRE].

[9] ATLAS collaboration, *Performance of the ATLAS Trigger System in 2015*, *Eur. Phys. J.* **C 77** (2017) 317 [arXiv:1611.09661] [INSPIRE].

[10] CMS collaboration, *The CMS trigger system*, 2017 *JINST* **12** P01020 [arXiv:1609.02366] [INSPIRE].

[11] D.P. Kingma and M. Welling, *Auto-encoding variational Bayes*, arXiv:1312.6114 [INSPIRE].

[12] J. An and S. Cho, *Variational autoencoder based anomaly detection using reconstruction probability*, *Special Lecture on IE* **2** (2015) 1.

[13] L. Lyons, *Open statistical issues in particle physics*, arXiv:0811.1663.

[14] E. Gross and O. Vitells, *Trial factors for the look elsewhere effect in high energy physics*, *Eur. Phys. J.* **C 70** (2010) 525 [arXiv:1005.1891] [INSPIRE].

[15] T.Q. Nguyen et al., *Topology classification with deep learning to improve real-time event selection at the LHC*, arXiv:1807.00083 [INSPIRE].

JHEP05(2019)036

[16] R.T. D'Agnolo and A. Wulzer, *Learning new physics from a machine*, *Phys. Rev.* **D 99** (2019) 015014 [arXiv:1806.02350] [INSPIRE].

[17] J.H. Collins, K. Howe and B. Nachman, *Anomaly detection for resonant new physics with machine learning*, *Phys. Rev. Lett.* **121** (2018) 241803 [arXiv:1805.02664] [INSPIRE].

[18] A. De Simone and T. Jacques, *Guiding new physics searches with unsupervised learning*, *Eur. Phys. J.* **C 79** (2019) 289 [arXiv:1807.06038] [INSPIRE].

[19] J. Hajer, Y.-Y. Li, T. Liu and H. Wang, *Novelty detection meets collider physics*, arXiv:1807.10261 [INSPIRE].

[20] A.A. Pol et al., *Detector monitoring with artificial neural networks at the CMS experiment at the CERN Large Hadron Collider*, *Comput. Softw. Big Sci.* **3** (2019) 3 [arXiv:1808.00911] [INSPIRE].

[21] CMS collaboration, *Anomaly detection using deep autoencoders for the assessment of the quality of the data acquired by the CMS experiment*, Technical Report, CERN, Geneva (2018).

[22] ATLAS collaboration, *Deep generative models for fast shower simulation in ATLAS*, ATL-SOFT-PUB-2018-001 (2018).

[23] T. Heimel, G. Kasieczka, T. Plehn and J.M. Thompson, *QCD or what?*, *SciPost Phys.* **6** (2019) 030 [arXiv:1808.08979] [INSPIRE].

[24] M. Farina, Y. Nakai and D. Shih, *Searching for new physics with deep autoencoders*, arXiv:1808.08992 [INSPIRE].

[25] B. Schölkopf et al., *Estimating the support of a high-dimensional distribution*, *Neural Comput.* **13** (2001) 1443.

[26] F.T. Liu, K.M. Ting and Z.-H. Zhou, *Isolation forest*, in 8$^{\text{th}}$ *IEEE International Conference on Data Mining (ICDM08)*, December 15–18, Pisa, Italy (2008).

[27] F.T. Liu, K.M. Ting and Z.-H. Zhou, *Isolation-based anomaly detection*, *ACM TKDD* **6** (2012) 3.

[28] C.C. Aggarwal, *Outlier analysis*, in *Data mining*, C.C. Aggarwal ed., Springer, Germany (2015).

[29] M. Gemici et al., *Generative temporal models with memory*, arXiv:1702.04649.

[30] T. Sjöstrand et al., *An Introduction to PYTHIA 8.2*, *Comput. Phys. Commun.* **191** (2015) 159 [arXiv:1410.3012] [INSPIRE].

[31] DELPHES 3 collaboration, *DELPHES 3, a modular framework for fast simulation of a generic collider experiment*, *JHEP* **02** (2014) 057 [arXiv:1307.6346] [INSPIRE].

[32] D. Contardo et al., *Technical proposal for the Phase-II upgrade of the CMS detector*, CERN-LHCC-2015-010 (2015).

[33] M. Cacciari, G.P. Salam and G. Soyez, *FastJet user manual*, *Eur. Phys. J.* **C 72** (2012) 1896 [arXiv:1111.6097] [INSPIRE].

[34] M. Cacciari, G.P. Salam and G. Soyez, *The anti-k_t jet clustering algorithm*, *JHEP* **04** (2008) 063 [arXiv:0802.1189] [INSPIRE].

[35] I. Higgins et al., *beta-vae: Learning basic visual concepts with a constrained variational framework*, (2017).

[36] J.M. Tomczak and M. Welling, *VAE with a vampprior*,arXiv:1705.07120.

JHEP05(2019)036

[37] F. Chollet et al., *Keras*, https://github.com/fchollet/keras, (2015).

[38] M. Abadi et al., *TensorFlow: Large-scale machine learning on heterogeneous systems*, (2015).

[39] D.P. Kingma and J. Ba, *Adam: a method for stochastic optimization*, arXiv:1412.6980 [INSPIRE].

[40] F. Pedregosa et al., *Scikit-learn: machine learning in Python*, *J. Mach. Learn. Res.* **12** (2011) 2825.

[41] CMS collaboration, *Search for pair production of third-generation leptoquarks and top squarks in pp collisions at* $\sqrt{s} = 7$ *TeV*, *Phys. Rev. Lett.* **110** (2013) 081801 [arXiv:1210.5629] [INSPIRE].

[42] CMS collaboration, *Search for third-generation scalar leptoquarks and heavy right-handed neutrinos in final states with two tau leptons and two jets in proton-proton collisions at* $\sqrt{s} = 13$ *TeV*, *JHEP* **07** (2017) 121 [arXiv:1703.03995] [INSPIRE].

[43] J. Duarte et al., *Fast inference of deep neural networks in FPGAs for particle physics*, 2018 *JINST* **13** P07027 [arXiv:1804.06913] [INSPIRE].

JHEP05(2019)036